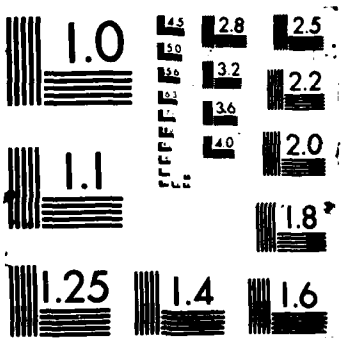
AD-A186 027 MAXIMUM LIKELIHOOD PRINCIPLE AND MODEL SELECTION WHEN THE TRUE MODEL IS U..(U) PITTSBURGH UNIV PA CENTER FOR MULTIVARIATE ANALYSIS R NISHII FEB 87 TR-87-01 1/1

UNCLASSIFIED AFOSR-TR-87-1017 F49620-85-C-0008 F/G 12/3 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0

DTIC FILE COPY

AD-A186 027

(When Data Entered)

ATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 87-1017	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): MAXIMUM LIKELIHOOD PRINCIPLE AND MODEL SELECTION WHEN THE TRUE MODEL IS UNSPECIFIED	5. TYPE OF REPORT & PERIOD COVERED: Journal - February 1987
6. PERFORMING ORG. REPORT NUMBER: 87-01	
7. AUTHOR(s): Ryuei Nishii	8. CONTRACT OR GRANT NUMBER(s): F49620-85-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Center for Multivariate Analysis, Fifth Floor Thackeray Hall, University of Pittsburgh, Pittsburgh, PA 15260	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304 A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research, Department of the Air Force, Bolling Air Force Base, DC 20332 NM	12. REPORT DATE: February 1987 13. NUMBER OF PAGES: 15
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): SCHMC CO 11	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DTIC ELECTE S OCT 16 1987 D

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

AIC, BIC, consistency, law of iterated logarithm

MLE, regularity conditions

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Suppose independent observations come from an unspecified distribution. Then we consider the maximum likelihood based on a specified parametric family by which we can approximate the true distribution well. We examine the asymptotic properties of the quasi-maximum likelihood estimate

DD FORM 1 JAN 73 1473

and of the quasi-maximum likelihood. These results will be applied to model selection problem.

AFOSR-TR- 87-1017

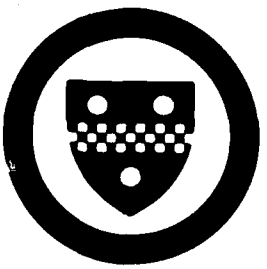
MAXIMUM LIKELIHOOD PRINCIPLE AND MODEL SELECTION WHEN THE TRUE MODEL IS UNSPECIFIED*

Ryuei Nishii

Center for Multivariate Analysis
University of Pittsburgh
and
Hiroshima University

Center for Multivariate Analysis

University of Pittsburgh



87 0 14

MAXIMUM LIKELIHOOD PRINCIPLE AND MODEL SELECTION WHEN THE TRUE MODEL IS UNSPECIFIED*

Ryuei Nishii

Center for Multivariate Analysis
University of Pittsburgh
and
Hiroshima University

DTIC
COPY
INSPECTED
6

February 1987

Technical Report No. 87-01

Accesion For
NTIS CRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or special
A-1

Center for Multivariate Analysis
University of Pittsburgh
Fifth Floor Thackeray Hall
Pittsburgh, PA 15260

* Research sponsored by the Air Force Office of Scientific Research (AFOSC), under Contract F49620-85-C-0008. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

MAXIMUM LIKELIHOOD PRINCIPLE AND MODEL SELECTION WHEN THE TRUE MODEL IS UNSPECIFIED*

Ryuei Nishii

Center for Multivariate Analysis
University of Pittsburgh

and

Hiroshima University

ABSTRACT

Suppose independent observations come from an unspecified distribution. Then we consider the maximum likelihood based on a specified parametric family by which we can approximate the true distribution well. We examine the asymptotic properties of the quasi-maximum likelihood estimate and of the quasi-maximum likelihood. These results will be applied to model selection problem.

AMS subject classification: Primary 62A10; secondary 62F12.

Key words and phrases: AIC, BIC, consistency, law of iterated logarithm MLE, regularity conditions.

* Research sponsored by the Air Force Office of Scientific Research (AFOSC), under Contract F49620-85-C-0008. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

1. INTRODUCTION

The maximum likelihood principle is the basic and useful technique for statistics. It has a long history and there is quite a bit of literature treating its asymptotic properties, e.g., Wald (1949) and LeCam (1953). These classical results are based on the assumption that the unknown density function lies in a specified parametric family. However, if this assumption is not true, do similar results remain valid? Cox (1961, 1962) considered first such a problem in testing of separated families, (see also Berk (1966, 1970)). Huber (1967) pointed out that this problem is connected with robust estimation. White (1982) reviewed this problem and showed the consistency and the asymptotic normality under the assumptions corresponding to the regularity conditions in the classical theory. Additional related references are Akaike (1973) and Foutz and Srivastava (1977).

In Section 2 we give the consistency order of the maximum likelihood estimator and of the maximum likelihood under the usual conditions with additional assumptions on higher order derivatives of the specified densities. Further we treat the testing problem of two families. Section 3 is concerned with the model selection. We prove the strong consistency of BIC type criteria in a very general setting. The inconsistency of AIC will also be shown. However, we reconsider the consistency in model selection in Section 4. All proofs of the theorems will be shown in Section 5.

2. OBSERVATIONS AND FAMILY OF DENSITIES

Let n observations (which may be multivariate) $x_1, \ldots, x_n$ ($e\ \mathbb{R}^d$) be independently and identically distributed as the probability density function g with respect to a fixed measure ν on $\mathbb{R}^d$. Suppose that $\int |\log g(x)| g(x) d\nu(x) < \infty$. Next consider the family of densities

$$M = \{f(x|\theta) | \theta \in \Theta\} \tag{2.1}$$

where Θ is a convex set in $\mathbb{R}^p$. Define the quasi-log-likelihood of n observations as

$$L_n(\theta) = n^{-1} \sum_{i=1}^{n} \ell(x_i|\theta), \qquad \ell(x|\theta) = \log f(x|\theta) \tag{2.2}$$

and define the quasi-maximum likelihood estimate by $\hat{\theta} = \hat{\theta}_n$. Recall the Kullback-Leibler information:

$$I(g;f,\theta) = \int g(x) \log\{g(x)/f(x|\theta)\} d\nu \geq 0 \tag{2.3}$$

provides some closeness from g to $f(\cdot|\theta)$. We call θ_g and $f(\cdot|\theta_g)$ the quasi-true parameter and the quasi-true density in M respectively when θ_g minimizes $I(g;f.\theta)$, $\theta \in \Theta$, or equivalently θ_g maximizes the expected log-likelihood

$$e(\theta) = \int g(x) \log f(x|\theta) dx. \tag{2.4}$$

Obviously if g(x) is exactly specified by M as $f(x|\theta_0)$, then $\theta_g = \theta_0$.

Now we make assumptions on (g,M) which will enable us to study the asymptotic behavior of maximum likelihood principle.

ASSUMPTION A1. The quasi-true parameter θ_g is unique and is an interior point of Θ.

ASSUMPTION A2. (a) $\ell_\alpha(x|\theta) = \partial\ell(x|\theta)/\partial\theta_\alpha$ and $\ell_{\alpha\beta}(x|\theta) = \partial^2\ell(x|\theta)/\partial\theta_\alpha\partial\theta_\beta$ $(\alpha,\beta = 1, \ldots, p)$ are measurable with respect to $x \in \mathbb{R}^d$ for each $\theta \in \Theta$ and continuous with respect to θ for each x, where $\ell(x|\theta) = \log f(x|\theta)$.

(b) $|\ell(x|\theta)|$, $|\ell_\alpha(x|\theta)|$, $|\ell_{\alpha\beta}(x|\theta)|$, $|\ell_\alpha f(x|\theta)\ell_\beta f(x|\theta)|$ are dominated by the integrable functions with respect to $g(x)$, which do not depend on θ.

ASSUMPTION A3. $V(\theta_g)$ and $W(\theta_g)$ are positive definite where

$$V(\theta) = E_g[\frac{\partial}{\partial\theta}\ell(x|\theta)\frac{\partial}{\partial\theta^T}\ell(x|\theta)] \text{ and } W(\theta) = -E_g[\frac{\partial^2}{\partial\theta\partial\theta^T}\ell(x|\theta)].$$

ASSUMPTION A4. There exists the quasi-maximum likelihood estimate of $\hat{\theta}_n$ which tends to θ_g with probability 1.

ASSUMPTION A5. (a) $\ell_{\alpha\beta\gamma}(x|\theta) = \partial^3\ell(x|\theta)/\partial\theta_\alpha\partial\theta_\beta\partial\theta_\gamma$, $(\alpha,\beta,\gamma = 1, \ldots, p)$ are measurable with respect to x for each θ.

(b) $|\ell(x|\theta)|^2$, $|\ell_{\alpha\beta\gamma}(x|\theta)|$ and $|\ell_{\alpha\beta}(x|\theta)|^2$ are dominated by integrable functions with respect to g, which do not depend on θ.

Remark on A4. (i) case $g \in M$: Several sufficient conditions ensuring the assumption A4 are known, e.g., Wald (1949), Huber (1967) and 5e.2 of Rao (1973).

(ii) case $g \notin M$: White (1982) showed that A1-A3 with A4': Θ is compact ensure A4. Conditions by Huber, derived without assuming that g is exactly specified, suffice A4. Also Wald's assumptions can be modified to this situation by substituting $df(x,\theta_0)$ for $g(x)d\nu$ and θ_0 for θ_g, which meet A4.

If the true density is completely unknown, any of our conditions is not checked. However, M gives a good approximation to g and M meets condi-

tions A1-A5 when $g(x) = f(x|\theta_0)$, then (g,M) will satisfy A1-A5.

The assumptions A1-A4 are corresponding to the regularity conditions in the classical theory. They ensure the strong consistency of $\hat{\theta}_n$ on $L_n(\hat{\theta})$. Further, the asymptotic normality of $\hat{\theta}_n$ can be shown, e.g., White (1982) and Foutz and Srivastava (1977). If we assume A5 additionally, the consistency order may be evaluated as in the following theorem which will play a key role in studying model selection criteria.

THEOREM 1. Let n independent observations come from the distribution with density g and (g,M) meet A1-A5 where M is defined in (2.1). The orders relating to the quasi-maximum likelihood estimates $\hat{\theta}_n$ and the log-likelihood are:

$$\text{(i)} \qquad \hat{\theta}_n = \theta_g + O((n^{-1}\log\log n)^{1/2}) \qquad \text{a.s.,}$$

$$\text{(ii)} \qquad L_n(\hat{\theta}) = L_n(\theta_g) + O(n^{-1}\log\log n) \qquad \text{a.s.,}$$

$$\text{(iii)} \qquad L_n(\hat{\theta}) = e(\theta_g) + O((n^{-1}\log\log n)^{1/2}) \qquad \text{a.s.,}$$

where θ_g is defined in A1, $L_n(\theta)$ in (2.2) and $e(\theta)$ in (2.4).

Note that Theorem 1 is new even if g is exactly specified by M. Under non-regular case the consistency order of $\hat{\theta}_n$ may be different from $O((n^{-1}\log\log n)^{1/2})$. However, (ii) still remains valid as long as the consistency order of $\hat{\theta}_n$ is faster than $O((n^{-1}\log\log n)^{1/2})$ because the order of (ii) is based on the law of iterated logarithm for $\ell(x_n|\theta) + \ldots + \ell(x_n|\theta_g)$.

Cox (1961, 1962) introduced the problem: Which family specifies the true density? He proposed the corrected likelihood ratio test. Our problem is: Which family is closer to the true density? We take a simple likelihood ratio approach. Let $M_i = \{f_i(x|\theta_i)|\theta_i \in \Theta_i\}$ $(i = 1, 2)$ be families of den-

sities (which may not be separated), and let ε_i be maximized expected log-likelihoods in M_i (see 2.4). Then test the hypothesis

$$H_0: \varepsilon_0 = \varepsilon_1 \quad \text{versus} \quad H_1: \varepsilon_0 > \varepsilon_1. \tag{2.5}$$

Assume both (g,M_i) satisfy A1-A5. If H_1 is true, from (iii) of Theorem 1 the likelihood ratio

$$\lambda_n = \sum_{j=1}^{n} \log\{f_0(x_j|\hat{\theta}_0)/f_1(x_j|\hat{\theta}_1)\} \tag{2.6}$$

tends to infinity since $n^{-1}\lambda_n \to \varepsilon_0 - \varepsilon_1 > 0$, a.s., which implies the likelihood ratio can asymptotically find the family closer to g. To make more detailed discussion, we get:

THEOREM 2. Consider the testing hypothesis (2.5) under the conditions A1-A5. Then the likelihood ratio test is consistent.

Let σ^2 be the asymptotic variance of $n^{-1/2}\lambda_n$. Then if $d \equiv |\varepsilon_0 - \varepsilon_1|/\sigma$ is large, we can discriminate the families by using small data. However, when d is small we need a large data. Hence in such a case it would be preferable to develop similar discussion as the corrected likelihood ratio proposed by Cox. See also Kent (1986).

3. MODEL SELECTION

We have shown that the likelihood ratio test is useful when two models are under consideration. When one has many models as the candidates for the true density g, model selection procedures are utilized. Consider k models $M_i = \{f_i(x|\theta_i)|\theta_i \in \Theta_i\}$. We treat here the criteria given by the following forms:

$$IC(i) = -2nL_n^{(i)}(\hat{\theta}_i) + c_n p_i, \quad (i = 1, \ldots, k) \tag{3.1}$$

where $\hat{\theta}_i$, $L_n^{(i)}(\hat{\theta}_i)$ and p_i are respectively the quasi-maximum likelihood estimate, the quasi-maximum log-likelihood divided by n and the number of parameters under the model M_i. The model minimizing (3.1) will be regarded as the best model. Akaike (1973) proposed to take $c_n \equiv 2$ (AIC), Schwarz (1978) and Rissanen (1978) proposed $c_n = \log n$ (BIC), and Hannan & Quinn (1979) as $c_n = K \log\log n (K > 0)$. Suppose the expected log-likelihood of M_1 is largest among those of k families. By Theorem 2, IC(i) $(i = 1, \ldots, k)$ will take almost surely its minimum value at IC(1) for large n if $\lim n^{-1}c_n = 0$. Every criterion above satisfies this condition. Hence we can find asymptotically which model is closest to g. Further we treat the case that the closest model M_1 (M; say) is divided into several subfamilies (nested case).

Suppose the quasi-true parameter vector θ_g can be written as

$$\theta_g = (\theta_1^*, \ldots, \theta_q^*, 0, \ldots, 0), \quad \theta_1^* \neq 0, \ldots, \theta_q^* \neq 0$$

and suppose zero vector is an interior point of Θ. This assumption implies that $\theta_{q+1}, \ldots, \theta_p$ are redundant. We call $J_* = \{1, \ldots, q\}$ the quasi-true model and $J_f = \{1, \ldots, p\}$ the full model for simplicity. Let J be a subset of J_f. Then submodel of M specified by J, say $M(J)$, is defined by $\{f(x|\theta(J))|\theta \in \Theta\}$ where $\theta(J) = (0\ 0\theta_{j_1} 0 \ldots 0\theta_{j_2} 0 \ldots 0\theta_{j_\ell} 0 \ldots 0)$, $J = \{j_1, \ldots, j_\ell\}$.

EXAMPLE. Let $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$, $g(x) = 1/2\{\phi(x-1)+\phi(x+1)\}$ and $\mu = \{\sigma^{-1}\phi(\sigma^{-1}(x-\mu))|\theta = (\theta_1,\theta_2) = (\sigma^2-1, \mu),\ \theta_1 > -1,\ -\infty < \mu < \infty\}$. Then $\theta_g = (1,0)$, $J_* = \{1\}$, $J_f = \{1,2\}$, $M(\{2\}) = \{N(\mu,1)\}$, $M(\{1\}) = \{N(0,\sigma^2)\}$.

Suppose $(g,M(J))$ meet A1-A5 and write the quasi-true parameter and the quasi-maximum likelihood estimate by θ_{Jg} and $\hat{\theta}_J$ respectively. Hence $e[\theta_{Jg}] = e[\theta_g]$ if $J \supseteq J_*$; and $< e[\theta_g]$ if $J \not\supseteq J_*$. Thus by Theorem 2;

THEOREM 3. Let λ_n be the likelihood ratio $L_n(\hat{\theta}_J) - L_n(\hat{\theta}_{J*})$. Then if $J \supseteq J_*$, $\lambda_n \geq 0$ and $\lambda_n = O(n^{-1}\log\log n)$, a.s. If $J \not\supseteq J_*$, $\lambda_n \to e(\theta_{Jg}) - e(\theta_g) < 0$.

THEOREM 4. Let $\hat{J}_n$ be a subset of J_f minimizing IC(J) of (3.1). If c_n satisfies both

$$\lim_{n\to\infty} n^{-1}c_n = 0 \quad \text{and} \quad \lim_{n\to\infty} c_n/\log\log n = +\infty, \tag{3.2}$$

then $\hat{J}_n$ is a strongly consistent estimator of the quasi-true model J_*, i.e., $\lim_{n\to\infty} \hat{J}_n = J$, a.s.

Note that if we relax the latter condition of (3.2) as

$$\lim_{n\to\infty} n^{-1}c_n = 0 \quad \text{and} \quad \lim_{n\to\infty} c_n = +\infty, \tag{3.3}$$

then $\hat{J}_n$ is a weakly consistent estimator of J_*, i.e., $\lim_{n\to\infty} P[\hat{J}_n = J_*] = 1$.

However, we need extensive calculation for getting $\hat{J}_n$ when p is large because there are 2^p-1 non-empty subsets of J_f. Our alternate procedure saves computation. Let $J_{-j} = \{1,\ldots,j-1,j+1,\ldots,p\}$ for $j \in J_f$. Define

$$\tilde{J}_n = \{j \in J_f | IC(J_{-j}) \geq IC(J_f)\}.$$

Then by the similar lines of the proof of Theorem 4, we get:

THEOREM 5. If c_n satisfies (3.2) or (3.3), then $\tilde{J}_n$ is also a strongly or weakly consistent estimator of J^*.

AIC is not consistent because $c_n \equiv 2$ does not meet (3.2) nor (3.3). It will overestimate the quasi-true model. The probability $\lim P[\hat{J}_{nAIC} = J] > 0$, for $J \supseteq J_*$ will be expressed using positive linear combinations of independent chi-square variates, however, its formula is hard to evaluate in a simple form.

4. DISCUSSION

Our results are based on the i.i.d. assumption. However, the Theorems 1-5 still remain valid even if n observations have weak dependency which ensures the central limit theorem and the law of iterated logarithm. Hence our results are quite general.

Next we try to reconsider the consistency in model selection problem. From the point of view that the model is an approximation with finite paramaters to the true density with infinite parameters (see Shibata (1980)), the quasi-true model under M becomes the full model in many cases. Then AIC also becomes consistent since it does not underestimate the quasi-true model. Our observations do not provide the difference of AIC and BIC in this case. Unfortunately our observations provide no difference of AIC and BIC in this case.

The purpose of the model selection may be to find the model by which we can get some good prediction for future observation, not the model which provides a good fitting for given observations. Recall AIC is proposed as an estimator of the *predictive* density. The consistency is one criterion for classifying the model selection procedures, and this criterion may not always lead a suitable conclusion in practical situation.

ACKNOWLEDGMENT

The author would like to express his thanks to Professors C.R. Rao and M.B. Rao for their comments.

5. APPENDIX

Proof of Theorem 1. From A1 and A4, $\hat{\theta}_n$ exists and is an interior point of Θ for large n. Employing Taylor's expansion we get

$$\underset{\sim}{Q} = \partial L_n(\hat{\theta})/\partial\theta = \partial L_n(\theta_g)/\partial\theta - W_n(\theta_g)(\hat{\theta}_n - \theta_g) + \underset{\sim}{r}_n \tag{5.1}$$

where

$$W_n(\theta) = -\partial^2 L_n(\theta)/\partial\theta\partial\theta^T : p \times p, \quad \underset{\sim}{r}_n = (r_{1n}, \ldots, r_{pn})^T,$$

$$r_{in} = (\hat{\theta}_n - \theta_g)^T[\partial^2(\frac{\partial}{\partial\theta_i} L_n(\bar{\theta}))/\partial\theta\partial\theta^T](\hat{\theta}_n - \theta_g),$$

$$\bar{\theta} = \theta_g + \varepsilon(\hat{\theta}_n - \theta_g), \quad 0 < \varepsilon < 1.$$

By the law of iterated logarithm and A3, A5, we have

$$L_n(\theta_g)/\partial\theta = O((n^{-1}\log\log n)^{1/2}), \qquad \text{a.s.}$$

$$W_n(\theta_g) = W(\theta_g) + O((n^{-1}\log\log n)^{1/2}), \quad \text{a.s.} \tag{5.2}$$

because $E\partial L_n(\theta_g)/\partial\theta = \partial E_g \ell(X|\theta_g)/\partial\theta = 0$. From (5.2) and A3, $W_n(\theta_g)$ is positive definite when n is large. Solving (5.1),

$$\hat{\theta}_n - \theta_g = W_n(\theta_g)^{-1}\{\partial L_n(\theta)/\partial\theta + \underset{\sim}{r}_n\}.$$

By A5 there exist H: an integrable function with respect to g and $K > 0$ such that for any $\alpha, \beta, \gamma = 1, \ldots, p$

$$|\partial^3 L_n(\bar{\theta})/\partial\theta_\alpha\partial\theta_\beta\partial\theta_\gamma| \le n^{-1}\sum_{i=1}^{n} H(x_i) < K,$$

which implies $\underset{\sim}{r}_n = O(1)(\hat{\theta}_n - \theta_g)$, a.s. Thus

$$\hat{\theta}_n - \theta_g = O((n^{-1}\log\log n)^{1/2}), \quad \text{a.s.}$$

Again by the law of iterated logarithm we know

$$L_n(\theta_g) = \mu_g + O((n^{-1}\log\log n)^{1/2}), \quad \text{a.s.}$$

where $\mu_g = e(\theta_g)$ is defined in (2.4). From (i) and A2,

$$L_n(\theta) - L_n(\hat{\theta}_g) = (\hat{\theta}_n - \theta_g)^T \partial L_n(\theta_g)/\partial\theta + 1/2(\hat{\theta}_n - \theta_g)^T \partial^2 L_n(\bar{\theta})/\partial\theta\partial\theta^T (\hat{\theta}_n - \theta_g)$$

$$= O(n^{-1}\log\log n), \quad \text{a.s.}$$

Hence, $L_n(\hat{\theta}) = L_n(\theta_g) + L_n(\hat{\theta}) - L_n(\theta_g) = O((n^{-1}\log\log n)^{1/2})$, a.s.

Proof of Theorem 2. The asymptotic normality of the likelihood ratio λ_n of (2.6) is known by Foutz and Srivastava (1977) as

$$n^{-1/2}\lambda_n \xrightarrow{L} N[\epsilon_0 - \epsilon_1, \sigma^2]$$

where $\sigma^2 = E_g[\log\{f_0(X|\theta_{0g})/f_1(X|\theta_{1g})\}]^2$ and θ_{ig} $(i = 0,1)$ are the quasi-true parameters. Using a consistent estimator of σ^2 as

$$\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}[\log\{f_0(x_i|\hat{\theta}_0)/f_1(x_i|\hat{\theta}_1)\}]^2,$$

we make the rejection region of H_0 by

$$R_n = \{\lambda_n > \sqrt{n}\,\hat{\sigma}_n\xi_\alpha\}$$

where ξ_α is the upper 100α-percent point of the standard normal distribution. Under H_1, or equivalently $\mu = \epsilon_0 - \epsilon_1 > 0$,

$$P[R_n^{(\alpha)}|H_1] = P[n^{-1/2}(\lambda_n - n\mu) \geq \hat{\sigma}_n\xi_n - n^{1/2}\mu|H_1] \to 1, \quad (n\to\infty)$$

because $n^{-1/2}(\lambda_n - n\mu) \xrightarrow{L} N[0,\sigma^2]$ and $\hat{\sigma}_n\xi_n - n^{1/2}\mu \to -\infty$ in P.

Proof of Theorem 4. If $J \not\supseteq J_*$, then

$$IC(J) - IC(J_*) = (\#J - q)c_n - 2n\{L_n(\hat{\theta}_J) - L_n(\hat{\theta}_{J*})\}$$

$$= \log\log n[(\#J - q)c_n/\log\log n - 2n(\log\log n)^{-1}\{L_n(\hat{\theta}_J) - L_n(\hat{\theta}_{J*})\}]$$

$$\to +\infty, \quad \text{a.s.} \quad \text{(Theorem 2)},$$

since $\#J - q > 0$ and $\lim_{n\to\infty} c_n/\log\log n = +\infty$. This implies for large n, $IC(J) > IC(J_*)$, a.s. Hence, $\hat{J}_n \subseteq J_*$.

If $J \not\subseteq J_*$,

$$IC(J) - IC(J_*) = 2n[L_n(\hat{\theta}_{J*}) - L_n(\hat{\theta}_J) - (\#J - q)c_n/(2n)] \to +\infty, \quad \text{a.s.}$$

since $L_n(\theta_{J*}) - L_n(\theta_J) \to c > 0$ and $\lim n^{-1}c_n = 0$. Hence, $\hat{J}_n \supseteq J_*$ for large n.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. Proc. 2nd Internat. Symp. on Information Theory, (Eds. B.N. Petrov and F. Czáki), Budapest: *Akademiai Kiado*, 267-81.

BERK, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* 37, 51-8.

BERK, R.H. (1970). Consistency a posteriori. *Ann. Math. Statist.* 41, 894-906.

COX, D.R. (1961). Tests of separate families of hypotheses. Proc. 4th Berkeley Symp. 1, 105-23.

COX, D.R. (1962). Further results on tests of separate families of hypotheses. *J.R. Statist. Soc. B*, 24, 406-24.

FOUTZ, R.V. and SRIVASTAVA, R.C. (1977). The performance of the likelihood ratio test when the model is incorrect. *Ann. Statist.* 5, 1183-94.

HANNAN, E.J. and QUINN, B.G. (1979). The determination of the order of an autoregression. *J.R. Statist. Soc. B*, 41, 190-5.

HUBER, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. 5th Berkeley Symp. 1, 221-33.

KENT, J.T. (1986). The underlying structure of nonnested hypothesis tests. *Biometrika*, 73, 333-43.

LeCAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Baye's estimates. Univ. California Publications in *Statist.* 1, 227-330.

RAO, C.R. (1973). Linear Statistical Inference and Its Applications. New York: Wiley.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465-71.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6, 461-4.

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* 8, 147-64.

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* 60, 595-601.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

END

12-87

DTIC